DOCUMENT RESUME

ED 284 871 TM 870 187

AUTHOR Van Epps, Pamela D.
TITLE Heuristics to Facilitate Understanding of Discriminant Analysis.
PUB DATE 30 Jan 87
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 29-31, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; Correlation; *Discriminant Analysis; Educational Research; *Heuristics; Mathematical Models; Multivariate Analysis; Predictive Measurement; *Predictor Variables; *Simulation; Statistical Studies
IDENTIFIERS Eigenvalues; Statistical Analysis System

ABSTRACT

This paper discusses the principles underlying discriminant analysis and constructs a simulated data set to illustrate its methods. Discriminant analysis is a multivariate technique for identifying the best combination of variables to maximally discriminate between groups. Discriminant functions are established on existing groups and used to predict the classification of subjects with unknown group membership. The simulated study sought to discriminate high, medium, and low job achievement groups comprised of 10 employees each. Five employee variables were created varying in their degree of correlation with the independent variable. A discriminant analysis was performed on this ideal data set with the CANDISC feature of the Statistical Analysis System (SAS). The rest of the paper treats in detail the statistical output of this analysis and serves as a guide to interpreting it. The discussion of results for the simulated data includes comparison of the univariate and multivariate statistics, use of the correlation matrix to detect multicollinearity, and application of the Wilks' Lambda test. The paper concluded with the graphing of centroids by discriminant functions as a means of visually identifying the different subgroups. (Author/LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED284871

Heuristics to Facilitate Understanding of Discriminant Analysis

Pamela D. Van Epps

University of New Orleans

New Orleans, LA 70148

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Van Epps

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX, January 30, 1987.

Tm 870187

Abstract

Heuristics, as a tool, serve to guide a researcher to discover for himself the theories behind a principle. This paper discusses the principles underlying discriminant analysis. Certain theories behind discriminant analysis are illustrated using concrete examples. Steps in the interpresentation process are delineated.

Introduction

Assume a social scientist has a group of subjects that he can distinguish based on a preplanned (a priori) number of factors or variables that constitute a score profile (Nunnally, 1978) for each subject. Assume also that the scientist wishes to analyze the two or more subgroups, where group membership is determined by the score profiles, to determine three things: 1. whether the subgroup's profile differences are statistically significant--keeping in mind, however, that statistical significance is primarily a function of group size (Carver, 1978); 2. whether the score profile variables can be combined in such a way that the discrimination among the subgroups can be maximized; and 3. whether any guidelines can be established by which new individuals can be placed into one of the subgroups. The multivariate statistical technique that the scientist would use to determine any or all three of these objectives is called discriminant analysis.

Basic Concepts

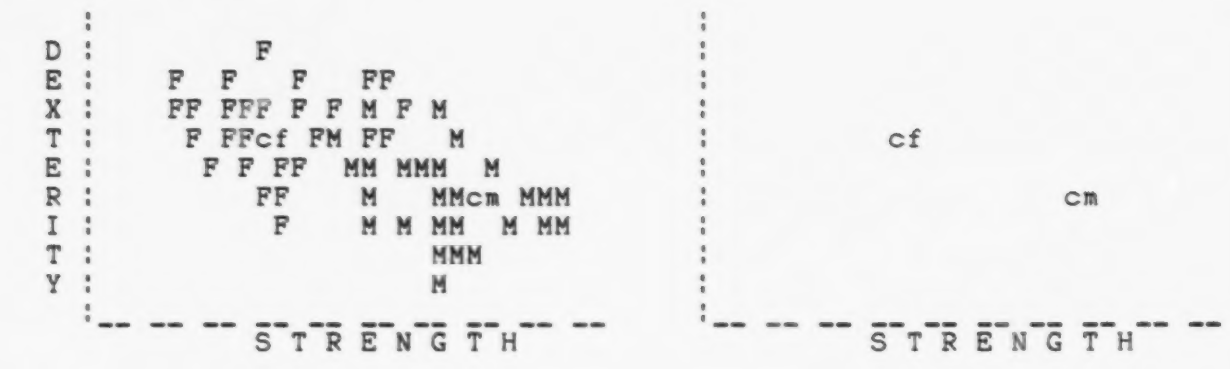
Because the research scientist wishes to statistically distinguish between two or more groups he first selects discriminating variables for measuring the characteristics, the

score profile, which he expects to differ across subgroups. The object is to create a linear combination of the discriminating variables so that the groups are as statistically distinct as possible. In other words, the researcher wishes to discriminate between groups, to tell them apart on a given dimension with one group clustering at one end, a second group at the other end, and perhaps a third group in the middle.

The technique can be visualized by plotting each subject on a graph in k-dimensional space where each axis represents one of the discriminating profiles which theoretically are orthogonal, or unrelated, to one another. Another way of visualizing the concept of discriminant analysis is to imagine a three dimensional box (assuming a three profile model) in which a number of subjects are clustered together based on their score profiles. One group would be clustered, perhaps, in one corner of the box while another group would be cluster in the middle, and a third group in another corner. The clustering must have little overlap if the discriminant analysis is to be useful for providing information.

A simplified example of discriminant analysis would begin with two groups, male and female, measured for two distinguishing characteristics--manual dexterity and manual strength. Because research supports the hypothesis that females have more hand dexterity than males and that males have more strength in their hands than females, for this example the researcher expects the

two groups to occupy two different regions of a 2-dimensional space with little overlap.



As the graph of the profile scores for each subject demonstrates, there is a clustering of the 27 males and 27 females that can be distinguished since each group tends to occupy a different region. As expected, the females have higher dexterity scores and lower strength scores while the males have lower dexterity scores with higher strength scores. There is also a noticeable degree of overlap between the groups.

An average of all the profile scores for each group on each variable is calculated (cf and cm) and is called a centroid. The centroids are the points about which the subjects in a group "center" in all directions. For groups that are well discriminated the centroids are far apart with each group's members hovering near their own centroid.

Since most studies utilize large numbers of subjects, instead of depicting a point for each profile score it is often more convenient to depict the region of group scatter by a circle

in which the centroid is designated. The overlap of the circles is the extent to which the two groups are not discriminated by the two variables.

Inherent in the discriminant analysis are concepts from the general linear model. Specifically, the discriminating variables are used to form one or more linear combination(s), called discriminant function(s), in the following form:

$$Y = a1X1 + a2X2 + aiXi$$

The Y, a synthetic independent variable, represents the score on the discriminant function, the a's are the weighting coefficients which are applied to the X's which represent the raw scores on the discriminating variables. For a given variable, the same weight is applied to each of the subject's scores in each group which results in a new score for each subject. Y simply combines the information from all the weighted X's.

The total number of unique and uncorrelated discriminant functions that can be derived in a given study equals the number of groups minus one or equals the number of discriminating variables--whichever number is less. As already noted, the object of the analysis is to form functions that maximize the separation of groups, between group variance, and functions that minimize the dispersion of scores within each group, within group variance (Reiff, 1986).

Study Background

In order to demonstrate the various aspects of discriminant analysis, data were created to represent the achievement level of 30 adult employees who had been given some expensive on-the-job training. Because the Director of Personnel does not wish to spend time training future employees who might not successfully translate the training into satisfactory job performance, he decided to analyze 30 employees who have already completed the training and to form discriminant equations. He selected 10 subjects who are considered to be high achievers (H), 10 middle achievers (M), and 10 low achievers (L). The Director also chose the following as discriminating variables: sex, raw score on a instrument designed to measure achievement, age, job satisfaction, and raw score on an instrument designed to measure a task score.

The hypothetical data were created for each of the discriminating variables in order to demonstrate the ability of variables to discriminate based on their degree of correlation with the synthetic independent variable. For example, sex was created so that all the High achievers were females and all the Low achievers were males with the Middle achievers about half and half. This discriminating variable should prove to be very strong. The raw scores for achievement were developed to show a moderately strong correlation with higher scores for the High

achievers and lower scores for the Low achievers. The age variable was developed to show a very slight relationship, with the older employees rating somewhat higher in achievement than the younger employees. The job satisfaction scores ranged from 1 to 5 and were arranged so that each number was used twice for both the High and Low achievers; high scores were assigned to the Middle achievers. The task score variable was intended to demonstrate a very weak discriminator and was therefore a simple listing of random numbers from a random number table. Each of the five discriminating variables demonstrates different levels of correlation with the synthetic dependent variable ranging from a very strong correlation to the anticipated very weak correlation withthe random numbers variable.

TABLE 1 - HYPOTHETICAL DATA

EMPLOYEE	ACHIEVES	SEX	RATING	AGE	SATISFAC	TASKSCOR
1	H	1	99	28	5	51
2	H	1	65	32	4	69
3	H	1	73	25	3	37
4	H	1	82	27	2	61
5	H	1	64	24	1	25
6	H	1	91	25	5	46
7	H	1	78	26	4	59
8	H	1	85	37	3	83
9	H	1	79	40	2	86
10	H	1	98	39	1	33
11	M	1	95	40	5	42
12	M	1	80	24	4	11
13	M	1	75	38	5	68
14	M	1	91	21	4	29
15	M	1	87	36	5	69
16	M	0	90	19	4	25
17	M	0	70	32	3	26
18	M	0	55	17	5	19
19	M	0	60	28	4	14
20	M	0	62	17	5	1
21	L	0	59	42	5	23
22	L	0	78	26	4	91
23	L	0	84	39	3	80
24	L	0	71	20	2	84
25	L	0	64	35	1	10
26	L	0	65	18	5	25
27	L	0	42	33	4	30
28	L	0	80	19	3	27
29	L	0	73	27	2	40
30	L	0	52	16	1	90

Study Results

The data were analyzed using the CANDISC feature of the SAS program for discriminant analysis (SAS Institute Inc., 1985). The first step included an examination of the univariate statistics to see if there were any interesting results. A look at the R-squared values in Table 2 shows that sex is the best explainer of the variance in this study, which is as expected. The second and third variables, rating and job satisfaction, were the next best at explaining variation. Interestingly, age was the least powerful variable, even more so than the random number variable, task score, which it turns out was not as random as expected.

TABLE 2

UNIVARIATE STATISTICS

VARIABLE	MEAN	TOTAL STD	WITHIN STD	BETWEEN STD
SEX	0.5	0.5085476	0.3042903	0.5
RATING	74.9	14.21352	13.26999	7.430343
SATISFAC	3.466667	1.407696	1.282359	0.8082904
AGE	28.33333	8.082904	8.24666	1.709776
TASKSCOR	45.13333	27.05516	25.70949	13.00205

AVERAGE R-SQUARED: UNWEIGHTED = 0.2545306

R-SQUARED	RSQ/(1-RSQ)	F	PROB > F
0.666667	2.000	27.000	0.0001
0.188472	0.232	3.135	0.0596
0.227378	0.294	3.973	0.0307
0.030859	0.032	0.430	0.6550
0.159278	0.189	2.558	0.0961

WEIGHTED BY VARIANCE = 0.1570554

The next step was the computing of a Correlation Matrix to confirm the correlations of the hypothetical data in order to

check for the possible problem of multicollinearity. If any two discriminating variables correlate too highly, the researcher would have to consider dropping one of them. Table 3 demonstrates that there is no problem with multicollinearity, therefore all five variables can be retained.

TABLE 3

TOTAL-SAMPLE CORRELATION MATRIX

VARIABLE	SEX	RATING	SATISFAC	AGE	TASKSCOR
SEX	1.0000	0.5653	0.0482	0.3104	0.2306
RATING	0.5653	1.0000	0.0869	0.2236	0.2104
SATISFAC	0.0482	0.0869	1.0000	-0.0232	-0.1837
AGE	0.3104	0.2236	-0.0232	1.0000	0.2431
TASKSCOR	0.2306	0.2104	-0.1837	0.2431	1.0000

Table 4 presents the results of the multivariate analysis. Of particular interest are the reported Canonical Correlations, the Squared Canonical Correlations, and the Eigenvalues for each of the two functions. [Remember the number of functions is either the number of variables or the number of groups less one, whichever number is lower. In this case 3 - 1 = 2 < 5.] For the first Canonical Correlation, 0.834, the squared value shows that the first function explains 69.5% of the variation. The second function explains an additional 33.6% of the variation between achievement classes. It should be noted that the first function Canonical Correlation is higher than the highest univariate R-squared statistic which supports the desirability of using a multivariate versus a univariate technique. Multivariate

analysis considers the results of all the variables and their interactions simultaneously as opposed to one at a time. This typically results in the most powerful prediction. Finally, the Eigenvalues which are the ratios of between class variance to within class variance for the two functions show the first function (2.28) doing the best job of discriminating; however, the second function (0.51) does have some unique predictive ability.

TABLE 4

	CANONICAL CORRELATION	ADJUSTED CANONICAL CORRELATION	APPROX STANDARD ERROR	SQUARED CANONICAL CORRELATION
1	0.833513	0.802714	0.056685	0.694744
2	0.579819	0.538538	0.123266	0.336190

EIGENVALUES OF INV(E)*H
= CANRSQ/(1-CANRSQ)

EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
2.2759	1.7695	0.8180	0.8180
0.5065	.	0.1820	1.0000

The next statistic to be examined is the Wilks' Lambda presented in Table 5. Because the closer this statistic approaches 0 the better is the discriminating ability of the functions, a value of 0.20 would be considered an excellent result.

TABLE 5

MULTIVARIATE TEST STATISTICS AND F APPROXIMATIONS
S=2 M=1 N=11

STATISTIC	VALUE	F	NUM DF	DEN DF	PR > F
WILKS' LAMBDA	0.2026322	5.619	10	46	0.0001
PILLAI'S TRACE	1.030934	5.106	10	48	0.0001
HOTELLING-LAWLEY TRACE	2.78239	6.121	10	44	0.0001
ROY'S GREATEST ROOT	2.275934	10.924	5	24	0.0001

NOTE: F STATISTIC FOR ROY'S GREATEST ROOT IS AN UPPER BOUND
F STATISTIC FOR WILKS' LAMBDA IS EXACT

From here the researcher is interested in determining which variables contributed most to the discriminating ability of the functions. In an actual research situation this would allow the researcher to drop the less effective discriminators. Table 6, the Between Canonical Structure values, describes the first function under column CAN1 and the second function under column CAN2. These are the correlation coefficients between each raw profile score variable and the synthetic discriminant score variable for the subjects. As expected sex and the achievement rating at .9983 and .9918 were the two best discriminators in the first function. The least interesting variable in the first function is age which was expected to be a very weak discriminator. In the second function job satisfaction and the random number variable, task score, did the best job at improving the explained variance between the groups after the effects of the first discriminant function were removed.

TABLE 6

TOTAL CANONICAL STRUCTURE

	CAN1	CAN2
SEX	0.9779	-0.0832
RATING	0.5166	0.0959
SATISFAC	0.0338	0.8210
AGE	0.1651	-0.1883
TASKSCOR	0.0641	-0.6821

BETWEEN CANONICAL STRUCTURE

	CAN1	CAN2
SEX	0.9983	-0.0591
RATING	0.9918	0.1281
SATISFAC	0.0591	0.9983
AGE	0.7835	-0.6214
TASKSCOR	0.1348	-0.9910

Finally, of particular interest are the centroids or Class Means, of the discriminant scores presented in Table 7. These are best understood, in a two dimensional case, by plotting them on a graph. Because the example yielded only two functions, it lends itself very well to this effort. The centroid values for H, L, and M show that the High achiever subgroup is most different from Low achiever subgroup. The plotting shows the centroids for the L and H groups are rather distinct and as expected are the furthest apart. The subjects for the M group are clustered half with the L group and half with the H group. The centroid for the M group therefore falls between the two groups. This tells the researcher that the discriminating variables are most effective at discriminating between the High achievers and Low achievers.

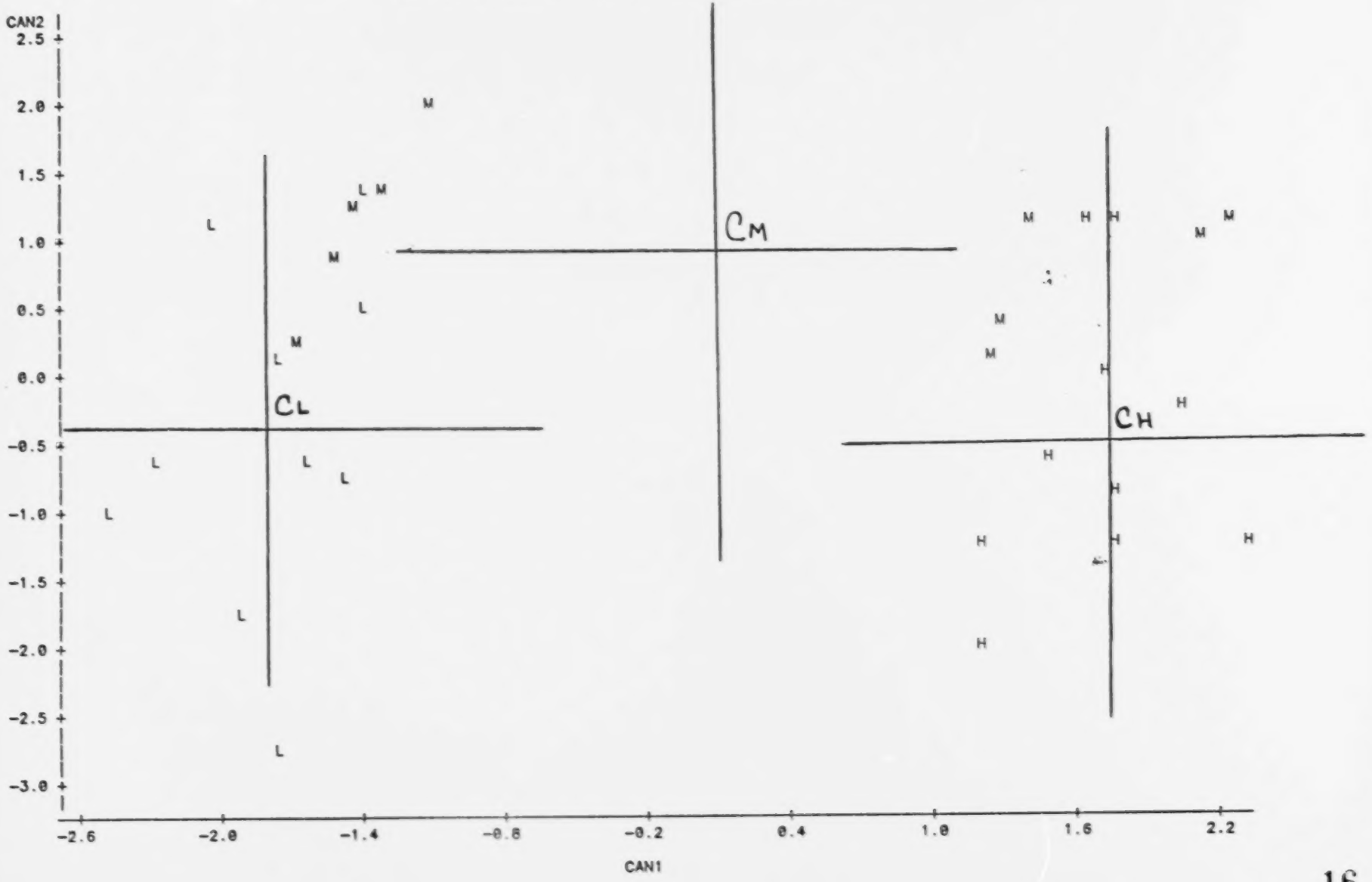
TABLE 7

CLASS MEANS ON CANONICAL VARIABLES

ACHIEVES	CAN1	CAN2
H	1.6900	-0.5254
L	-1.8096	-0.4277
M	0.1196	0.9531

PLOT OF CANNONICAL DISCRIMINANT FUNCTIONS

PLOT OF CAN2*CAN1 SYMBOL IS VALUE OF ACHIEVES



SUMMARY

Discriminant analysis, is a multivariate technique which offers the researcher a unique set of steps by which he can measure the success with which discriminating variables can actually discriminate when combined into discriminant functions or he can predict the classification of subjects with unknown group membership. This paper presented some substantive examples to help explain the basic concepts and illustrates one application. Some detail focused on the results of selecting various discriminating variables based on their degree of correlation with the independent variable. Some of the statistical results of the illustration discussed were the comparison of the univariate and multivariate statistics, multicollinearity, and application of the Wilks' Lambda test. The paper concluded with the graphing of centroids by discriminant functions as a means of visually identifying the different subgroups (Huck et al., 1978).

REFERENCES

Carver, R.P. (1978). "The case against statistical significance testing." Paper presented at the annual meeting of the American Educational Research Association, New York.

Huck, S.W., Cormier, W.H. & Bounds, W.G. (1974). Reading Statistics and Research. New York: Harper & Row.

Nunnally, J.C. (1978). Psychometric Theory. New York: McGraw-Hill, Inc.

Reiff, Henry B. (1986) "Heuristics for Demonstrating Discriminant Function Analysis: A Monte Carlo Example." Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis.

SAS Institute Inc. (1985). SAS User's Guide: Statistics, version 5 Edition. Cary, NC:SAS Institute Inc.